DOCUMENT RESUME

ED 053 662 HE 002 401

AUTHOR Harmon, Lindsey R.
TITLE Fourteen Years of Research on Fellowship Selection; A Summary.
INSTITUTION National Academy of Sciences - National Research Council, Washington, D.C.
REPORT NO Publ-1420
PUB DATE 20 Jul 66
NOTE 44p.

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS *City Government, *Educational Needs, Government Employees, *Higher Education, Inservice Education, *Public Administration Education, *State Government
IDENTIFIERS *Connecticut

ABSTRACT

This report, prepared for the New England Board of Higher Education (NEBHE), is one of six concerned with the public service training and education needs of the New England states. Section I discusses the background and tradition of public service training in Connecticut. Section II deals with in-service training programs provided by state and local agencies, by the Institute of Public Service at the University of Connecticut, and by other institutions. The third section examines the basic philosophy and theory underlying the structure of personnel training and education in the State, and the response of educational institutions to the State's manpower needs. Section IV discusses current characteristics of public service training in Connecticut, including the need to improve the image of governmental service and the whole training and education effort. Section V deals with the need for public service training and education in state and local government, including the need for top-level commitment to the realization of executive level training. Section VI examines the relationship between present and future training policies and the effectiveness and implementation of public policy; and Section VII considers 5 alternatives for meeting the requirements of public service education. (AF)

FOURTEEN YEARS OF RESEARCH ON FELLOWSHIP SELECTION

A SUMMARY BY LINDSEY R. HARMON

FOURTEEN YEARS OF RESEARCH ON FELLOWSHIP SELECTION

A SUMMARY BY LINDSEY R. HARMON

PROGRESS REPORT ON RESEARCH
SUPPORTED BY THE
NATIONAL SCIENCE FOUNDATION

JULY 20, 1966

OFFICE OF SCIENTIFIC PERSONNEL
NATIONAL ACADEMY OF SCIENCES
NATIONAL RESEARCH COUNCIL

NATIONAL ACADEMY OF SCIENCES

NATIONAL RESEARCH COUNCIL

Publication 1420

1966

Available from

Printing and Publishing Office
National Academy of Sciences
2101 Constitution Avenue
Washington, D.C. 20418

Price: $1.50

Library of Congress Card Catalog Number 66-61932

PREFACE

The Office of Scientific Personnel of the National Academy of Sciences—National Research Council has, from its inception during World War II, been broadly concerned with all aspects of the production and utilization of scientific manpower. The Division of Research, initiated within the Office of Scientific Personnel in 1952, has been concerned with data on the baccalaureate and doctorate origins of PhD's in all fields, with particular emphasis on the sciences. It has produced a series of publications relating to doctorate production, and the backgrounds and careers of PhD's. One of its major tasks over the period 1952-1966 has been the pursuit of excellence in fellowship selection—research devoted to the constant improvement of the techniques and procedures by which candidates for fellowships are to be evaluated. This work has been supported throughout by the National Science Foundation, toward whose predoctoral fellowship programs the research has been oriented.

Throughout its history, the Office of Scientific Personnel has operated under the direction of Dr. M. H. Trytten. The first Director of Research was Dr. Calvin W. Taylor, who took 2 years leave of absence from the psychology department of the University of Utah to get the research program under way. He was succeeded in 1954 by Dr. L. R. Harmon, who carried on this program as well as other manpower studies, and who has had the chief responsibility for the research program during its most active period. Dr. Harmon worked closely with Dr. Claude J. Lapp who directed the Fellowship Office during this time. Mr. Herbert Soldz, Data Processing Manager of OSP, who joined the Office of Scientific Personnel staff as research associate in 1956, wrote two of the Technical Reports the results of which are summarized in the present document. In 1960, Dr. John A. Creager joined the research unit, and over the next 6 years carried out a series of projects described in Technical Reports 18 through 26.

Throughout this period, the research program has had the benefit of review and suggestions by an advisory committee of senior scientists appointed by the President of the National Academy of Sciences. While in no way responsible for the results of the research, this committee, through its wise counsel, has been of great benefit to the research unit. It is hoped that the results of the selection research carried out over this 14-year period by these many hands will be informative and useful to the academic community as a whole, as well as to government agencies concerned with the support of graduate education.

CONTENTS

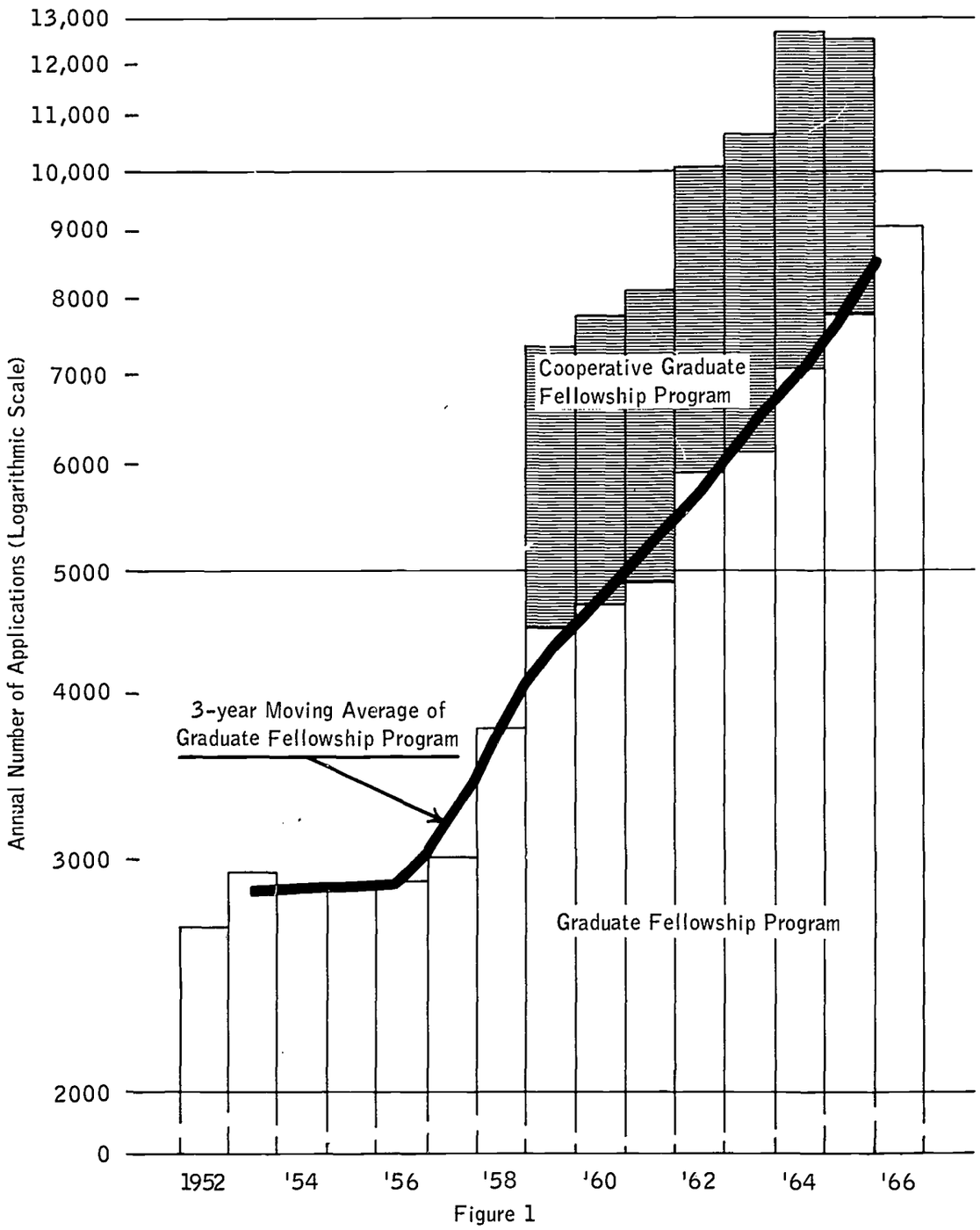


Figure 1

GROWTH IN APPLICATIONS FOR GRADUATE FELLOWSHIPS, 1952 THROUGH 1966
(Dated for Year in Which Tenure Would Begin)

INTRODUCTION

From the inception of the National Science Foundation Graduate Fellowship Program in 1952, through 1956 there was but slight growth in the number of applicants for these fellowships; the program was essentially stable in size. But beginning in the fall of 1956—fully a year before Sputnik I—there began a growth in number of applicants for NSF graduate fellowships that has continued to the present time. The fall of 1957 saw a spurt in the growth rate, and the size of the applicant population has grown, with minor irregularities, at the rate of 12 percent per annum from 1956 through 1966. On top of this growth in the basic Graduate Fellowship Program, there was initiated in 1959 the Cooperative Graduate Program, which, before it was terminated, pushed the total graduate applicant load to over 12,700 per year. Figure 1 displays these growth trends. The Cooperative Program was then phased out in favor of the Traineeship Program, which does not involve the central selection problems of its predecessors but puts the burden of selection on the graduate departments themselves. The program of research in fellowship selection has operated over this whole period; it is testimony to the farsightedness of the founders of the program that research designed to improve selection was initiated at the beginning. In this way, the fruits of the research have been steadily fed back into the program itself. One result, shown in the latest technical report, is that, within the limits of the information available at the time of selection, it would be difficult indeed to improve on the selections that have been made (26, p. 41).

This report of the findings of 14 years of research in fellowship selection techniques will, in the main, be organized around several research topics rather than chronologically. Yet, because the research necessarily proceeded stepwise from the simpler and more immediate problems to those more recondite and of a long-term nature, the report will also have a chronological arrangement. It will be a review primarily of the research findings; the changes in panel procedures and data processing which were introduced either on the basis of research findings or through less formalized experience will be given briefer treatment.

The first problem to receive research attention was the satisfactoriness of the operational instruments—i.e., application forms, grade report sheets, reference reports, and other associated documents. Two of the earliest technical reports are concerned with the tabulation and organization of the comments received from panel members with regard to improvement in these operating forms (1, 5, 28). Next, attention was turned to the reliability of the instruments: to what extent could one depend on the stability of the readings they gave? Some improvements were effected, and reliability was increased where necessary. Attention was then turned to the operations of the panels and use of research data for making the panel deliberations as efficient and effective as possible, thus conserving panelists' time and insuring that all feasible provisions were made for their arriving at the best possible decisions. As soon as sufficient time had elapsed to make evaluations of the selections in terms of later on-the-job performance, research was extended in this direction. In the most recent period, this is the aspect of the work which has received primary research attention.

SELECTION BY ABILITY ONLY

The aim of the application evaluation process carried out by committees of the National Academy of Sciences—National

Research Council is to insure that appointments are made on the basis of ability only. This requirement, written into the basic law by the Congress, is left to the good judgment of the program administrators to define, and they in turn have left the determinations of ability to the judgment of the evaluation panels within broad administrative guidelines. The research program which has been carried out over these 14 years has done much to provide a picture of what the abilities are that equip the graduate students for effective study and later functioning as scientists. Research has shown that, in an operational sense, the ability involved has remained remarkably constant. This is, no doubt, in large part due to the manner in which the evaluation process has been performed. Before beginning a description of research findings, it is therefore desirable to sketch, however briefly, the operational context of the evaluation process and its attendant research program.

APPLICATIONS PROCESSED

Each fall, all of the graduate departments in fields in which support is available and all of the colleges and universities that might send students on to graduate school in these fields are informed of the NSF Fellowship Program and are provided with the necessary forms for their students to use in the first step in applying for fellowships. The completed forms required for application are received and processed by the staff of the Fellowship Office of the National Academy of Sciences–National Research Council. Any matters that are legal or financial in nature are reviewed by the National Science Foundation Fellowship Section. Candidates for graduate fellowships are required to take the Graduate Record Examinations (if taken in a previous year, these scores may be used), and the results of these examinations are forwarded to the Fellowship Office for incorporation into the individual applicant records. The necessary forms and test scores are assembled into a folder for each applicant, and these folders, together with appropriate rosters, are ready for the evaluation panels when they arrive to make their judgments, typically in the middle of February.

PANELS BY FIELD OF SPECIALIZATION

In the Graduate Fellowship Program, applications in a given discipline are evaluated by a panel of experts within that general field. Typically, the panelists are senior professors in the nation's colleges and universities, but they include also scientists and engineers from industry and from nonprofit organizations. All panelists serve by appointment of the President of the NAS-NRC. At present, there are panels in the Graduate Fellowship Program in each of eight areas: anthropology and psychology; biological sciences; chemistry; earth sciences; engineering; mathematics; physics and astronomy; and social sciences. Each panel is presided over by a chairman who has previously served as a panelist. The average turnover from year to year on the panels is about one third, so that the majority of panelists at any one time have had at least one year's previous experience on a panel. In this way expertise which can be gained only from practical experience is preserved and passed on from year to year, but the rate of turnover is sufficient so that new people and new ideas are continually appearing and being given a chance to operate.

ROSTERS, LEVELS, AND QUALITY GROUPS

In addition to the folders which contain individual information, each panel is pro-

vided a set of rosters with the names of all the fellowship applicants within its field. These rosters are divided by applicant level so that some subpanels evaluate terminal candidates (those within one year of the PhD), others evaluate first-year candidates (mostly seniors who will enter graduate school the following fall), while still other subpanels evaluate the intermediate candidates (all other levels of graduate study). Each panelist is given a roster with the names of a set of candidates and the condensed record of the quantitative evidence regarding each candidate—his test scores, over-all ratings on reference reports, undergraduate science grade point average, and a "Summary Score" (a machine-computed score which is a weighted composite of these quantitative evidences of ability). Typically, each record is read independently by two panel members; these independent evaluations are then compared. If they are in essential agreement, the evaluations (which are performed on a ten-step scale) are averaged for the final score for that individual. If the two separate evaluations are not in essential agreement, the record may be read by a third, or even a fourth evaluator before a final score is computed. These final scores are then put in rank order by the panel chairman, and the whole list divided into "Quality Groups." Quality Group 1 is the most able group of candidates, comprising a certain percentage of the applicants in that field. The groups proceed downward in order to Group 6, which is defined by the requirement that none of the panel members would accept the candidate as a graduate student in his department. An attempt is made to have each group be as nearly homogeneous in ability as is feasible with such a limited number of groups. Members of Group 1 have all been offered fellowships; members of Group 2 have usually been offered fellowships; members of Group 3 may be offered fellowships, depending on the number of openings available and other factors to be described later. Those in Group 3 who do not get fellowships are given "honorable mention" if they request it, which most do. Occasionally, members of Group 4 may be given "honorable mention." Group 5 members are considered below the "honorable mention" level, but still capable of effective graduate school work. As described above, members of Group 6 are deemed unsatisfactory as graduate students. As a rule, Group 6 is small; the self-selection among applicants for fellowships is such that the candidates as a whole are far above the level of the general population of graduate students in these fields.

GEOGRAPHIC AND FIELD FACTORS APPLIED

With the determination of Quality Groups, the work of the evaluation panels is completed. The Quality Group results are then turned over to the National Science Foundation for the selection of awardees and the administration of the fellowships for those who accept. The National Science Foundation also determines, from among those in Group 3, who will be offered fellowships. This is done primarily on the basis of geographic factors, with a view to redressing imbalances which may have occurred because a disproportionate number of members of Groups 1 and 2 may have come from some geographic regions, whereas other regions may have had few applicants in Quality Groups 1 and 2. As a secondary consideration, field of specialization may also be taken into account in making awards from Quality Group 3. Perhaps it should be mentioned here that in the early days of the program, before the number of fellowships offered was as large as it has

been in recent years, it was only Quality Group 1 which was automatically offered fellowships; Group 2 was the one on which the geographic and field factors were brought into play.

VARIATIONS BY TIME AND PROGRAM

The preceding paragraphs describe the procedures typical of those used in the evaluation of the new applicants for NSF graduate fellowships. Holders of fellowships who seek renewals of their fellowships have, since 1956, been separately considered. Prior to 1956, renewals filed applications all over again in competition with the new applicants—usually including the repeating of the Graduate Record Examinations. That procedure was changed, and new forms were devised, seeking to ascertain whether the performance during the fellowship year was such as to warrant renewal—a somewhat different process from consideration of a candidate de novo. The procedure was somewhat different also for candidates for the cooperative fellowships. In that program, all the candidates from a given institution, regardless of field, were considered by one panel, and the panels as well as the applicant groups were interdisciplinary. Institutions rank-ordered their recommended candidates, and these rank orders were given careful consideration by the panels, being changed only when it appeared to the evaluation panel that an error had occurred. The panels then had the problem of splicing together the lists from the various schools in such a way as to arrange the candidates in Quality Groups that were as comparable as possible to those of the Graduate Program.

APPLIED RESEARCH AND DEVELOPMENT

This, in condensed form, was the context within which the research program operated. It sought to improve the evaluation procedures by examining as objectively as possible their reliability and validity and recommending changes when evidence suggested that improvements could be made. It was thus an applied research program aiming at practical usable results, rather than one of basic inquiry into all the interesting facets of the selection process. In the course of the 14 years under review here, 26 Technical Reports were issued on various aspects of the program. In addition, there were a number of Special Reports designed for limited distribution to those most immediately concerned with the operation of the program, and several working papers and memoranda relating to operational problems. This report will focus attention chiefly on the research findings of the Technical Reports but will include the other papers where appropriate. Throughout this report, references will be made to the basic research documents that are listed in the appended bibliography.

AN ADVISORY COMMITTEE

From the inception of the fellowship selection research effort, its work has had the advantage of overview by a research advisory committee. In the earliest days, this committee was composed primarily of research psychologists, chosen for their experience with problems of selection, either of students or of other personnel,

or for their expertise in research design. Later, the committee was broadened to include members of other disciplines, and its responsibilities were also broadened to include research on scientific manpower problems other than those directly connected with fellowship selection. Throughout the years of this research program, this committee has furnished invaluable help and guidance with regard to problems to be attacked and techniques which might be employed. In no wise, however, are the members of this committee responsible for research results or the failure to undertake research which might have been done; responsibility for the conduct of the research program rests directly on the staff of the Office of Scientific Personnel, operating within the boundaries determined by the financial support provided by the National Science Foundation.

RELIABILITY OF INSTRUMENTS AND PROCEDURES

A basic requirement for any evaluation tool is reliability—that is, reproducibility of results. Without knowledge of the degree of reliability of the various measuring instruments at its disposal, a panel must depend far too much on guesswork with regard to the weight to assign to each "instrument reading" or score. It was inevitable, then, that an examination of the reliability of the evaluation instruments, and of the evaluation process as a whole, was one of the first and most basic tasks of the research program.

GRE HIGHLY RELIABLE

The reliability of the Graduate Record Examinations was not examined by the Office of Scientific Personnel; results of this nature have been provided by the Educational Testing Service, the test publisher. Scores on the Verbal, Quantitative, and Advanced-field Tests are of a high order of reliability. These tests have been carefully designed with regard to a suitable balance of length, item difficulty, and content to maximize the results for a given amount of testing time. Typical reliability coefficients for the GRE are .90 for the Verbal Test, .84 for the Quantitative Test, and .82 to .94 for the various Advanced Tests. This is well above the minimum customarily required for tests used in individual selection (25, p. 7, referring to Graduate Record Examination Scores for Basic Reference Groups, Third Printing, Educational Testing Service, Princeton, N. J., 1961).

GPA SATISFACTORY

The undergraduate science grade point average (GPA) cannot really be evaluated in terms of reproducibility, as the students cannot be put through the same courses a second time. It is possible to determine a random-halves kind of reliability by splitting the courses taken by each individual on a random basis into two equivalent sets, and correlating the grade point averages computed separately for the two sets. This has not been done, and it is therefore necessary to depend on less direct evidence. Estimates by other researchers, using various methods, have indicated reliabilities for grade point averages upwards from .65.* Prediction studies, where grades are predicted by tests which are themselves not completely reliable, have shown test validities ranging into the .60's and .70's. Such validity coefficients require that the GPA would have substantially higher reliability, probably in the .80's or higher. It has therefore not seemed necessary to make a redetermination on the fellowship applicant population.

REDUNDANCY HELPS REFERENCE REPORTS

The reliability of the reference reports submitted in support of a fellowship application has come under careful study. The method used here is to compare the degree of agreement of one reference reporter, taken at random, with another, similarly selected at random, with regard to a given applicant. The average degree of ageement for a large number of applicants then describes the reliability or reproducibility of the ratings submitted by the reference reporters. It has been found that the agreement between two random raters

*A minimum figure estimated for graduate school grades, in Interpreting College Grade Averages, by Lewis B. Ward, Educational Testing Service, Princeton, N. J., April 1958.

is expressed by a correlation coefficient of about .31 to .36 (30, p. 10). Although this is a relatively poor agreement, it is not unexpected in view of all the sources of variation when one individual is evaluated by another. The reliability of the individual ratings can be substantially increased, however, by pooling the results from two or more raters. A statistical technique (the Spearman-Brown Prophecy Formula) is available to estimate the increase of reliability which is attained by adding any given number of raters. In practice, it is usually not feasible to have more than three or four raters—seldom are more than that many professors adequately acquainted with the qualifications of a given candidate. However, if the ratings by three knowledgeable raters are combined, a quite respectable reliability is attained, described by a correlation coefficient of .63. If four reporters with equivalent knowledge are available, their average will have a reliability of .69 (31, p. 10). These figures express the coefficients that would be expected if the average of several raters were to be correlated with the average of several other raters of equal knowledgeability. These results are illustrated graphically in Figure 2. With this understanding of the reliability of the reference report average rating, further statistical studies employing this variable could be undertaken with a known degree of confidence.

PANELISTS AGREE QUITE WELL

If the raters who submit reference reports are not in close agreement, what about the panel members themselves? How well do they agree with each other in their assessment of the evidence available to them? From the standpoint of reliability (but not validity) they are in a more favorable position to agree than are the reference reporters, for all of them have exactly the same information available upon which to base their judgments. The reference reporters, on the other hand, have known the candidates in a variety of different contexts, and so would necessarily base their ratings on somewhat different sets of factors. As it turns out, in comparing one panel member with another, the degree of agreement between any two of them is almost exactly equal to the degree of agreement between two sets of three reference reporters—about .70. When the composite evaluation of three panel members is taken, the reliability (as compared with a theoretical comparable set of three) is .87—a very satisfactory figure, indeed, to use as a measure of the stability of a consensus judgment upon which Quality Group is based (31, p. 11).

RE-EVALUATION OF REAPPLICANTS

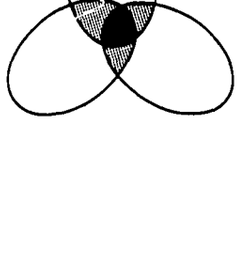
In the early days of the program, when every application was considered as a new application even though it came from a current Fellow, independent determinations of Quality Group were made in 2 successive years on the same people for a substantial number of cases. In these instances, of course, different sets of evidence were used in the 2 years—even the undergraduate grades being different for those who applied once in their senior years (before all grades were available) and again while in graduate school. For 396 cases who applied in 1952 and 1953, the year-to-year Quality Group judgment was found to agree to the extent of a contingency coefficient of .53, where the maximum coefficient possible would be .82 because there were only six categories of Quality Group available (27, p. 1). This finding was shortly rendered obsolete

A

The area of agreement of two raters whose ratings correlate .36 is shown as the overlap of these two ovals. The sum of these two ratings would correlate .53 with the sum of another similar pair.

B

This diagram illustrates the three-way overlap of three independent ratings, each of which correlates .36 with each of the others. The agreement of the sum of these three with the sum of another similar triad is expressed by a correlation coefficient of .63.



C

Here the correlation of .36 between any two raters taken at random builds up to a reliability coefficient of .69 if the sums of two independent sets of four ratings are correlated. The diagram illustrates the overlapping by pairs, by triads, and finally the common ground for all four raters.

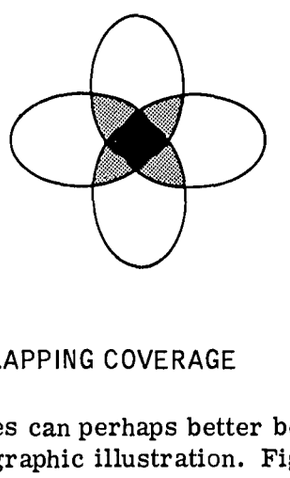


Figure 2

RELIABILITY OF RATINGS AS OVERLAPPING COVERAGE

because of a change in procedure for the renewal cases, but it may serve as a useful point of reference for the degree of agreement to be expected under circumstances such as those that obtained in 1952, 1953, and 1954. For these same 3 years, the mean over-all rating on the reference reports submitted on reapplicants was found to correlate .55 from one year to the next (27, p. 5).

The significance of these reliability figures can perhaps better be appreciated by a graphic illustration. Figure 3 serves to illustrate the matter. Let us assume that the large circle in Figure 3 represents perfect reliability—that there is a complete overlap between one reading and another. Any deviation is represented by a diminished overlap or coverage. The smaller concentric circles represent varying degrees of reliability. Each of these inner circles is tagged with a correlation

coefficient to denote the size of the reliability correlation illustrated. When this coefficient has declined to .7, only about half of the large circle is covered—the square of .7, or 49 percent. Such circle diagrams serve to illustrate accurately the mathematical significance of any coefficient of correlation, but in some cases the nature of the relationship is such as to make this kind of illustration seem strained. We will have occasion to use it again, however, when we come to the matter of

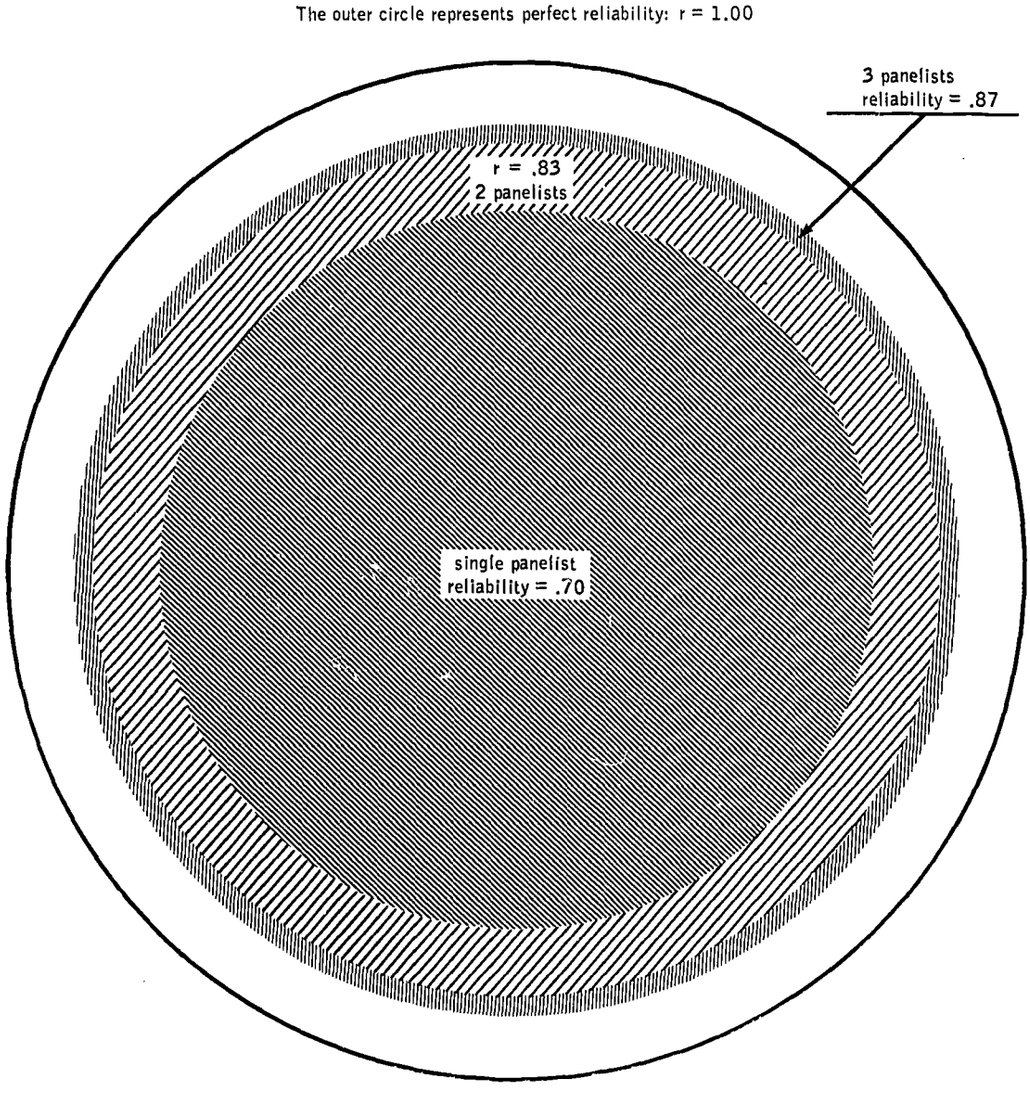


Figure 3

RELIABILITY OF PANEL EVALUATION AVERAGES
DIAGRAMMED AS APPROACHES TO COMPLETE RELIABILITY

validity—the extent to which some criterion of later accomplishment is correlated with, or represented in, the predictor instruments.

ARE ON-THE-JOB CRITERIA RELIABLE?

As the program advanced from consideration of the selection instruments available for panel use to the validity of these instruments as predictors of later on-the-job effectiveness, the question of the reliability of the on-the-job measures arose. Here there were several achievement measures to be considered. The most useful are the Confidential Reports of Performance and the citation counts. For various groups of cases, evaluated separately by field of specialization, the degree of inter-reporter agreement was found in the first study (of AEC applicants of 1949) to vary from .30 to .66 (15, p. 5). As there were multiple raters, the composite ratings had reliabilities from .66 to .87—quite comparable to the evaluations made by the NAS-NRC panels. In later studies considerable care was taken with respect to the selection of the Confidential Reports to be included in criterion composites; only those were retained in which the reporters had had a substantial opportunity to observe pertinent behavior. It should be expected, therefore, that the average of the Confidential Report ratings remained a reliable measure in the later studies, although no specific redeterminations of reliability were made.

CITATION RELIABILITY GOOD, AND IMPROVING

The citation counts derived from the Science Citation Index have proved to be a valuable addition to a composite measure of effective on-the-job functioning. As yet only fragmentary evidence is available, as there has not been sufficient time for the cases recently studied to have acquired substantial bibliographies which might be cited. Nevertheless, the citation counts derived independently from the 1961 and 1964 Science Citation Indexes intercorrelated about .45 to .60 for various groups of 1955-1956 first-year and intermediate fellowship applicants, separated by field (26, p. 62). It is quite reasonable to expect that, with publication over a longer period of time, and with more years of the Index to be averaged, these counts would prove to be even more reliable than the average of several Confidential Reports.

AN INFORMAL ESTIMATE

When several measures are combined, the reliability of the composite is greater than that of the elements taken separately. Such a composite of several elements was developed in the study of the terminal candidates of 1952-1954 to establish a criterion against which the validity of the various predictors could be checked. An informal estimate was made of the reliability of the composite selected for use in this study. This composite included various elements: the several ratings in each Confidential Report form (separately weighted); several indices of excellence derived from spontaneous comments in the Confidential Report forms; number of publications; income from scientific or technical work; and an index based on the number of times the individual was nominated as an outstanding younger scientist by other members of the follow-up group. These various elements were weighted to produce a single criterion composite score for each individual. This composite criterion was estimated to have a reliability of .70 to

.85, although there is no way in which a precise measure of such reliability can be obtained directly (23, p. 12). With the addition of citation counts, the reliability might be increased slightly, and the validity increased substantially.

This summarizes, in brief fashion, the information available regarding the reliability of the selection instruments, the operations of the panels, and the evaluations made of on-the-job functioning several years after fellowship application. In all cases, the measures available have sufficient reproducibility that attention can well be turned to other aspects, such as the nature of these variables and the extent to which the early measures may validly be used to predict the later ones.

AIDS TO THE DETERMINATION OF QUALITY GROUP

The essential function of the evaluation panels is the determination, for each candidate, of a Quality Group, or judged level of ability. A substantial amount of evidence is accumulated regarding each candidate; the panel members are required to examine this evidence and render a judgment. Anyone who has observed the panels at work can testify to the great amount of sincere effort and deep thought that is given to these decisions. Frequently, when there is disagreement among panelists, each shred of evidence is given repeated scrutiny and sometimes is debated with the agonizing uncertainty of a jury trying to reach a decision on an important trial. As the number of applications increased, as indicated by the logarithmic growth curve in Figure 1, this burden of work mounted steadily. It became imperative to do whatever could be done to marshall the evidence in such a way as to minimize the time the panelists needed to obtain a clear picture of the individual data in each case, and to render the best possible decisions with consistency and fairness. A considerable amount of developmental effort was therefore put into this aspect of the program, utilizing the research evidence that had been accumulated, and deriving further evidence as necessary.

FIRST CRUDE BEGINNINGS

The first research evidence regarding the correlates of panel judgment came from examination of the work of the 1952 evaluation panels. A very crude score was computed for each case and correlated with the Quality Group decision. This score was simply the count of the number of variables (V, Q, A, and reference report average) in which the individual scored above the 20th percentile. Each person thus earned a score of 0, 1, 2, 3, or 4, depending on whether he exceeded this very low cut point on one or more of the predictors. The resulting correlation was substantial enough to encourage further study (29, p. 1). In 1953 a somewhat finer scale was used. The range of scores on each of these four predictors was divided into five segments, or quintiles, by the 20, 40, 60, and 80 percentile points. No points were awarded for scores in the lowest quintile, one point for scores in the second quintile, etc., up to 4 points for scores in the highest 20 percent range. With four variables, this resulted in a possible range of scores from zero to 16. This simple scale correlated .69 with panel judgment, indicating that this work was probably worthy of further development (29, p. 2).

FROM THE SIMPLE TO THE METICULOUS

A further refinement of scales was made in 1955: instead of quintiles, the percentile scores were used. For each person, the sum of his percentile ranks on each of the four variables was computed, and this sum, with a range of zero to 400, was correlated with Quality Group. Each field and level was studied separately; the correlations ranged from .59 to .87, averaging .82 for Level 1, .75 for Level 2, and .70 for Level 3 (30, p. 6). It was apparent from these studies that the panelists were basing their decisions heavily, but not exclusively, on these items of quantitative evidence. It followed that it should be possible to express this evidence in a more succinct fashion, perhaps weighting it to correspond with the typical weights operationally (though not necessarily with full awareness) assigned to these variables by the panels in making their judgments. This led to the first experimental development of the Summary Score, which was not put into

operational use until after an experiment to determine its effect on panel judgment.

PROCEDURAL ADVANCES

While this experimental work was going on, other steps were being taken to organize and present the data in better fashion for the panel members. In the first 2 years of the National Science Foundation Fellowship Program the panel members had only the Scaled Scores on the Graduate Record Examinations. It was decided that it would be more meaningful if the NSF percentile ranks of these scores were presented on the rosters. This was done, and conversion tables to Scaled Scores were made available (29, p. 2). In subsequent years various other changes were made. For example, the reference report over-all ratings were presented, not as the average of all raters, but individually for each rater, to show at a glance the degree of agreement or disagreement among the reference reporters. Whatever the form of the test scores, GPA, and reference reports, however, they remained a discrete series of variables to be evaluated. The evidence from statistical studies was that the panelists went through some process of summation of these scores, either implicitly or explicitly, to get a general view of the level of ability they indicated, and also examined the individual reference reports and other application materials to get a more detailed and individualized view of each candidate. It was reasoned that if this first rough summation process could be performed by machine, a considerable amount of time could be saved—time which was spent by the panelist in a relatively routine-level arithmetical computation. One of the items which required a great deal of time to scrutinize and summarize was the Grade Report Sheet, itself a summarization in more standardized form of the evidence of the transcript, which is quite unstandardized. It was decided after an experimental "dry run" that it was feasible to have the applicants themselves compute a grade point average that would be checked by the Fellowship Office staff before being placed on the roster. This additional item presented on the roster was also available for inclusion in a generalized Summary Score, for which research work was under way.

EVOLUTION OF THE SUMMARY SCORE

Research done during 1955 indicated that, at the first-year level, the GRE scores were given heavy weight by the panels in determining Quality Group, whereas, at the intermediate and terminal levels, the reference reports and other evidence, such as the plan of research, were given heavier weight (30, p. 5). It was found that there were also field variations in the weights which had to be assigned to the various predictors to produce the maximum correlation of a composite of these scores and the Quality Group decision of the judges (30, p. 6). Yet these variations were not large, and could very well be averaged, in the interests of simplicity, to produce a single weighted composite. In effect, this composite came close to taking the average weight for each variable from a series of regression equations based on the various fields and levels. It might be noted here that the regression equation takes into account not only the correlation of each variable with the criterion (Quality Group in this case), but also the intercorrelations of the variables with each other. The weighted composite of these predictors was called the Summary Score—a term that has remained in use ever since (31, p. 15). This Summary Score was intended

to function as a simulation of the mental calculations performed by the average panelist in his subjective weighting of the five predictor scores (tests, GPA, and reference report average) in determining Quality Group.

AN EXPERIMENTAL PANEL

In 1956, an experimental panel of five experienced panel members was assembled and asked to evaluate a sample of fellowship candidates in an operation parallel to, but independent of the actual panels. These panelists were given the Summary Scores for the candidates, but asked to take a challenging, rather than an acquiescent view of this score—to examine the evidence for reasons to "disagree" with the Summary Score. It was intended, in this way, to minimize the effect of the score, which, it was felt, might otherwise overweight the quantitative evidence as compared with the other evidence in the files. The experimental panel was also given a "control" group of cases in which the Summary Score was not present. The judgments of this experimental panel (which was interdisciplinary and considered a set of cases cutting across all fields) were then examined to see whether the provision of the Summary Score made a difference in the judgmental process. The evidence, under these experimental circumstances, was that it did make a difference; the panelists made judgments more in line with the Summary Score when they knew what it was than when they did not (31, p. 16). Some doubt was cast on the extent to which generalizations could be made from these findings, however, because of the fact that the panels and applicants were both interdisciplinary, and because all three levels were considered by the panel. In other words, as the panelists read cases across all fields and levels, with only a few cases at each field-level combination, it seems probable that they would depend more heavily on the quantitative evidence than they would if they could concentrate on a single field and level with a larger number of cases.

SUMMARY SCORE TO THE RESCUE

By 1959 the panel work load had become very great. That was the year in which the Cooperative Graduate Program was introduced, necessitating the recruitment of a whole extra set of panel members. The possible use of the Summary Score had been discussed with panel chairmen in the intervening years, and it was decided to try it out on an experimental basis, with provision for careful statistical examination of the results, to determine whether it was having an undue "pulling" effect on Quality Group judgments. An exhaustive analysis indicated that there was no appreciable effect on the final judgments, but it was estimated that substantial time was saved by helping the panelists to arrive more quickly at the same decisions they would have made more laboriously if the score had not been provided (16, p. 13). A research program was recommended to re-examine periodically the operational weights assigned to the several quantitative predictors in the Summary Score to insure that it remain optimal for its objective of simulating the summarizing process of the panel members.

WEIGHTS REDETERMINED

The development and reweighting of the Summary Score was done on the data derived from the 1960 and 1961 graduate fellowship applications. Each field and

level was examined separately. The particular objective of this effort was to determine whether it would be necessary or advantageous to employ a specially weighted Summary Score for each different field-and-level combination. It turned out (19, p. 11) that, although some statistical increment to the validity of the Summary Score would result from such individuation, it would not be substantial and would detract from some of the present uses of the Summary Score where comparisons of quality across fields or across levels are desired. On the average, it was found that the Summary Score predicted Quality Group with a correlation of about .85. This correlation was highest with the first-year students, where the quantitative evidence had been found from the earliest studies to have the greatest weight, and least with the terminal applicants, where the plan of research is an important element among the items of evidence (19, p. 8). The plan of research is absent or given little weight at the first-year level and is available in only a portion of the intermediate cases. Undergraduate grades, in the form of the GPA, carry less weight with the panel members as graduate grades and reference reports become available. Among the GRE scores, it was found that the Advanced Test carried most weight in predicting Quality Group, and the Verbal Test least; in fact, it had less weight than did the GPA, which followed the Quantitative Test in relative significance as a predictor. The reference reports were given substantial weight at all levels, but predominant weight at the terminal level (19, p. 21).

SAME WEIGHTS FOR COOPERATIVES?

A similar investigation with respect to the prediction of Quality Group was undertaken with the Cooperative Graduate applicants. It was found that, although there were some specific differences between the two programs, very similar weights served for optimal prediction composites in both graduate programs. This is equivalent to saying that, from the standpoint of the panel evaluations, ability is defined in closely similar ways in the two programs (21, p. 15). Although it would be possible to make minor increments to the validity of the Summary Score as a predictor of final Quality Group by using special composites, there are also hazards attendant upon such use. These problems of complexity and possible error, and real difficulties in communication—particularly when it becomes necessary, as it occasionally does, to transfer a case from one panel to another—resulted in a decision to use only a single Summary Score formula, but to update this formula by periodic re-examination of the empirical weights. Such re-examinations have shown no substantial change over the past decade, indicating a de facto constancy in the definition of "ability" as judged by the evaluation panels. A check in 1965 showed a continued high correlation of Summary Score and Quality Group.

LISTS A AND B

The panel work load continued to climb at a rate averaging 12 percent per year, as indicated in the introduction to this report. This has necessitated a continuing search for means to economize on the time of the panelists, as it is not feasible to continue increasing the panel size and time indefinitely at the rate of 12 percent per year. One of the ways which grew out of the successful use of the Summary Score was a division of the applicants into two groups. It was found that none of the people with

a Summary Score below a certain point achieved awards, and only very few received honorable mention (33). It is obvious that the functional distinctions in the evaluation program are made above this cutting score—the separation into Quality Groups which may result in a fellowship offer or honorable mention. It is at this level that the most meticulous attention must be paid to the evaluation of evidence. The applications of individuals with Summary Scores below this point are nevertheless examined to determine whether there is any evidence which would result in a higher Quality Group than that predicted by the Summary Score. Occasionally such cases are found, and they are then evaluated in the same fashion as those in the primary list. This separation of the applicants into the two lists results in a substantial saving of time that can then be devoted to the crucial decisions, those between Quality Groups 2, 3, and 4.

THE PREDICTION OF DOCTORATE ATTAINMENT

From the standpoint of the evaluation panels, the proximate problem is to determine which of the fellowship applicants are most likely to succeed in graduate school. The achievement of success as a scientist following graduation is important, but it is a more distant goal, and much less likely to be attained if one is unable to complete the doctorate degree. The prediction of doctorate attainment is thus an intermediate step, and the validity of such prediction is one test of the effectiveness of the selection process. This criterion of success is also readily measured. It is available on all candidates through the Doctorate Records file of the Office of Scientific Personnel and requires less time to "mature" than do on-the-job criterion measures. Attention was therefore focussed on the extent to which each of the predictor variables available to the fellowship panels could predict doctorate attainment, and on the validity of a composite of such predictors.

A COMPOUND EFFECT

The earliest study using doctorate attainment as a criterion was made with the 1949 AEC fellowship applicants. It was found that in this group, separated into awardees and nonawardees, and further, into the biological and physical sciences, no satisfactory prediction could be made of the rate at which the doctorate degree was attained. It was found, however, that awardees attained the degree more than one year earlier than did nonawardees, a result of both higher ability and the advantage of holding a fellowship (31, p. 18). In this study, ability was controlled only by allowing for differences on the Verbal and Quantitative Tests. Research has since shown that these were the less significant tests to control. A more adequate study, done with larger numbers of cases, employed all three GRE scores as controls.

SEPARATED INTO ITS COMPONENTS

The next study employed the 1952 NSF fellowship applicant group, eliminating those for whom the ability tests were not available. The conclusion of this study was that those who received awards were more likely to complete the doctorate, and to complete it earlier, than those who did not obtain awards. The difference in time to completion found here was approximately a year. About 4 months of this time span were attributed to differences in tested ability, and more than 7 months were attributed to the award itself—or to the award and other abilities not measured by the test scores (18, p. 30).

A comprehensive investigation of the prediction of doctorate attainment was not made until 1965, however. One study focussed on the predictive significance of the GRE, using a specially selected sample of fellowship applicants designed to simulate, as nearly as possible from the available data, the range of talent among all science students applying for graduate school entrance. This required heavy selection from among the less able of the NSF fellowship applicants, as these applicants are above the general level of ability of graduate students as a whole. A second study used the 1955-1956 fellowship candidates. Both studies contributed significant results.

A SAMPLE TO SIMULATE A POPULATION

The study of the simulated "general graduate student" group indicated that the

Graduate Record Examinations are valid predictors of doctorate attainment, whether the criterion be percentage of people attaining degrees or time required for their completion. The validity of the tests varies, however, the Advanced Test being the best predictor in all fields. Poorest prediction was found in biology (25, p. 21). All predictions are modest in this simulated sample, ranging from a low of .06 for a group of female biology students to .55 for a group of female chemistry students. Part of the variability is a matter of chance fluctuations with small samples. Typical validities are in the .20's and .30's for the Verbal and Quantitative aptitude tests; validities occasionally rise into the .40's for the Advanced Test, and range up to .50 for a composite made up of all three tests taken together in an optimally weighted composite (25, p. 27).

MORE VARIABLES INCLUDED

With the follow-up sample of 1955 and 1956 candidates, the GPA and reference report averages were studied, as well as the GRE. One of the findings was that, although the Advanced Test was decidedly the best predictor, the composite of three tests was a better predictor than the Advanced Test alone. The GPA added a bit to the validity of the three-test composite; the reference report average made no further contribution as far as prediction of doctorate attainment is concerned. Separate composites were derived, by regression analysis, for first-year and intermediate candidates in various fields, for men and women separately. However, when the validity of these composites was compared with that of the Summary Score very little advantage was found. The special composite validities ranged from .17 to .55, while Summary Score validity ranged from .15 to .46 (25, p. 37). There is a minor amount of inflation in the special composite validites because of capitalization on chance errors; if this inflation is removed, the coefficients are even closer to those of the Summary Score. In the same study, the ability of the evaluation panels to predict doctorate attainment (using Quality Group as a predictor) was evaluated. It is found that, in spite of the crudity of the six-step Quality Group scale, it yields validity coefficients that are only slightly lower than those of the Summary Score. Quality Group coefficients ranged from .18 to .44 for the same set of field-level-sex groups. An incidental finding was that, holding constant field of specialization and Advanced Test score, the doctorate attainment rate of the women was lower than that of the men.

PREDICTING ON-THE-JOB EFFECTIVENESS

Doctorate attainment constitutes a clear-cut and proximal goal of the selection process. But it is on-the-job functioning which determines the final effectiveness of a program of fellowship support. Are the people who were supported more effective as scientists and engineers than those who were not supported? If so, is this superiority correlated with the ability measures used in selecting the fellows, or is it due solely to the fact of support itself? If initial ability correlates with the later on-the-job criteria, which measures are the best predictors, and what is the best method of combining and using the predictors to maximize the precision of the selection? These are some of the questions which the long-range research program on fellowship selection sought to answer. At this stage, the answers are only partial, but the results are definitely encouraging from the standpoint of the validity of the work of the evaluation panels.

SOME VALIDITY, SOME PUZZLES

In the earliest study of on-the-job effectiveness (the follow-up of Atomic Energy Commission fellowship candidates) it was found that prediction was possible within the awardee group but not within the non-awardee group. In the awardee sample positive validity was found for the GRE tests, the GPA, and reference report scores. The composite of these variables (using weights which approximated those of the Summary Score) had validity coefficients ranging from .28 to .66 (15, p. 8). These groups were small and quite possibly not representative of the larger numbers of candidates who came later to the NSF Program. In any case, further studies employing the more substantial numbers of NSF fellowship candidates were undertaken as soon as sufficient time had elapsed for these later candidates to graduate and establish themselves in jobs.

MORE SUBSTANTIAL SAMPLES

The terminal candidates, of course, have the shortest time lapse between fellowship application and occupational functioning, so it was the terminal-level candidates of the first 3 years of the NSF Program who were first studied. There were 2,178 members of this candidate population of 1952, 1953, and 1954 combined. All were included in the study, but it was possible to obtain satisfactory on-the-job criterion data on only about two thirds of this group (22, Figure 1). A number of criterion elements were collected, including data from questionnaires filled out by the former candidates themselves, and Confidential Reports completed by people whom they nominated as knowing most about their scientific and technical accomplishments. These former candidates were also asked to name three or four of the outstanding younger men in their general fields. A portion of those so named were themselves members of the follow-up group; the number of nominations received in this manner became another criterion element. An additional and very useful element was added later: citation counts from the Science Citation Index.

VALIDITY A COMPLEX MATTER

To check on the stability of the validity determinations, two samples very nearly equivalent in qualifications were used: the 1952 terminal-level candidates as one group, and the 1953 and 1954 terminal candidates combined as the other group. These groups were again sorted by employer category in one set of computations, and by field in another, to determine

whether the various job situations and fields of specialization had a significant relationship to the validity of the selection instruments. Each of the four quantitative measures of initial ability (tests and reference reports) was individually tested for its validity in predicting each of ten criterion elements. The ten elements included peer nominations; income; number of publications and patents; rank; a "fellowship recommendation" item; spontaneous comments under the headings "faint praise," "immature," "positive personality traits," "excellence;" and an over-all rating on scientific or technical contributions. Quality Group was treated as a predictor also, in spite of the limitation on its validity imposed by the rather coarse grouping. This produced several hundred validity coefficients. Their main significance can be shown rather succinctly. The most generally useful and meaningful criterion element, the average of the Confidential Report ratings of on-the-job performance, was predicted rather well by the earlier reference reports. The reference reports also predicted peer nominations and a number of indices of excellence derived from spontaneous comments on the Confidential Reports. Income is slightly but positively predicted by the reference reports, and rather well predicted by the Quantitative Test, when all fields are combined. When fields are separated, this prediction breaks down; apparently it is an artifact due to the circumstance that the higher-paying fields are also those in which quantitative capacity is most important (engineering and physics on the one hand, versus biology and psychology on the other) (22, pp. 10-11).

COMPOSITE VALIDITIES ARE HIGHER

A considerable improvement in validity is made when the several tests and the reference reports are combined, by a multiple-correlation technique, into a single score used to predict the various criterion elements. All correlations become positive, whether considered by field, by employer category, by award status, or by year of application. The best of the criterion elements, the over-all rating, is predicted with validity coefficients ranging from .23 to .43—certainly modest, but within the range ordinarily expected of such amorphous criteria as are afforded by measures of occupational effectiveness (22, p. 15). There were evidences that it would be possible to build up a more reliable and predictable criterion. This could be done by combining the various spontaneous Confidential Report citations for excellence in teaching, publications, research, and administration, the several different rating scales, and the peer nominations. This was not done in this study, however, as the main focus was not on maximum correlations, but on determining the relative contributions of the various predictors to measures of on-the-job performance. It was discovered that the relative weights shown by the regression equations were similar to those used in the Summary Score—a second validation of the general operational procedure used by the evaluation panels. This was confirmed by the positive correlations of the Quality Group decisions with the job effectiveness criteria (22, pp. 10-11).

TEACHERS MUST COMMUNICATE

An additional finding worthy of note in this study of the early terminal candidates could afford a useful point of departure if it should ever become important to predict teaching ability per se. This was a correlation of .25 between citations for excellence in teaching and ratings on communications ability in the reference reports submitted at the time of application.

While this correlation is modest, it is rather high for that of a single specific rating scale with citations many years later. About half of the people in academic work received one or more citations for excellence in teaching.

MORE CRITERION ELEMENTS

A much more detailed study of criterion measures was made in a follow-up of the first-year and intermediate-level candidates of 1955 and 1956. In these groups, there was an additional predictor not present in the previous studies—the undergraduate grade point average. More importantly, there was a significant new criterion element—the count of citations derived from the Science Citation Index (26, Appendix 3). There was a departure, too, in technique, in that the various criterion elements were built up into composites which had greater stability than any of the individual elements. Two criterion composites, somewhat different in conception, were initially developed—one designed to be maximally predictable by means of the fellowship application data, and the other designed to be maximally relevant as a measure of contributions to scientific advance. It was found, interestingly enough, that the composite of maximum relevance was only slightly less predictable than was the composite designed specifically for predictability (26, p. 32). The final composite which was used in most of the findings was the "most relevant" composite with the addition of doctorate attainment. This was referred to as the "inclusive" criterion.

As in the preceding studies, composites of predictors were formed to optimally predict the various criteria. It was found, however, that these composites were but slightly different from the Summary Score. The weights designed to be optimal in the Summary Score for prediction of Quality Group were also very nearly optimal for predicting on-the-job performance several years later (26, p. 31). Quality Group itself, used as a predictor, suffers only a small additional loss in validity, chiefly because of the fact that it is expressed only on a six-step scale, losing the variance which exists within each of the Quality Groups. Individual predictors were found to have relationships with criterion elements in a pattern which is quite understandable: the reference report ratings are optimal for prediction of on-the-job criteria (chiefly ratings) while the best predictor of doctorate achievement is the Advanced Test score (26, p. 33). The validites of two of the best predictors, taken individually, are shown in Figure 4. The Advanced Test of the Graduate Record Examination and the average of the over-all ratings on the reference reports are diagrammed. These diagrams do not show the overlap of these two predictors, but only their relative validities for each of four criterion elements: doctorate attainment, a count of number of publications, a count of number of citations in the 1964 Science Citation Index, and the over-all rating of the Confidential Reports of on-the-job effectiveness. Figure 5 shows the validity of the Summary Score as a predictor of the "inclusive" criterion composite. In this diagram, the contributions of all five of the predictors and their interactions with the complex of elements included in the criterion composite are lumped in the single figure, the validity coefficient of .39. This is the average of the validities in several different groups of cases.

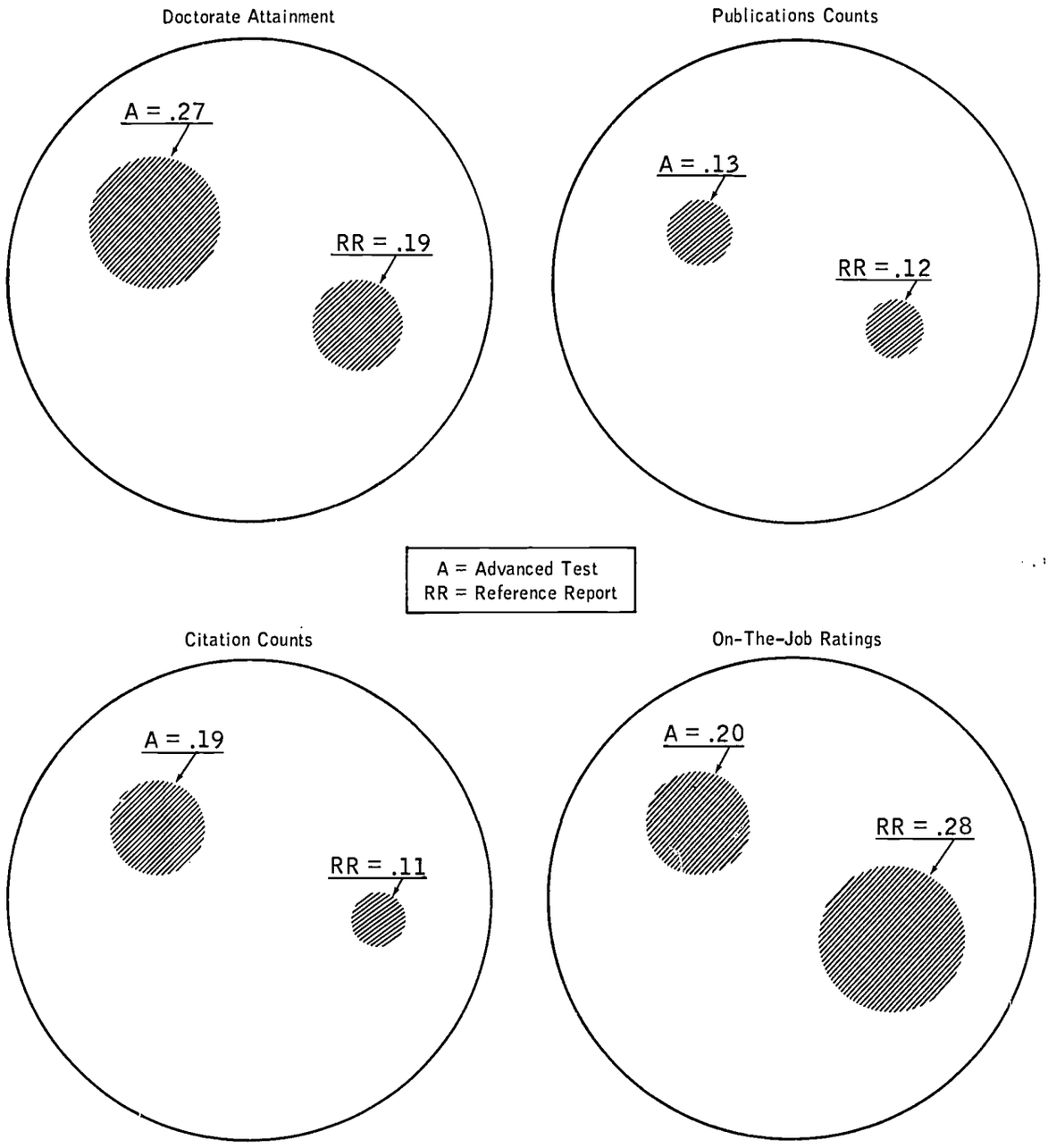
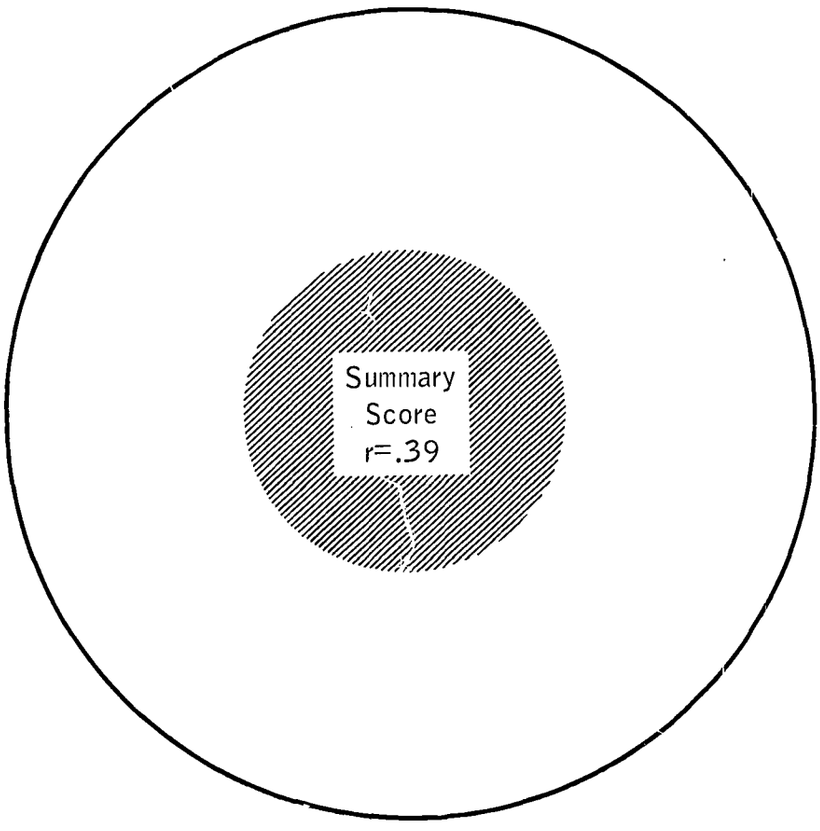


Figure 4

DIAGRAMS OF THE VALIDITY OF TWO PREDICTORS OF FOUR CRITERION ELEMENTS



23

Figure 5

VALIDITY OF THE SUMMARY SCORE AS PREDICTOR OF THE "INCLUSIVE" CRITERION

CANDIDATE DIFFERENCES BY FIELD, BY REGION, AND BY LATER EMPLOYER CATEGORY

One technical report focussed attention on differences in ability patterns in the various fields of specialization. In this research program, however, very little attention was paid directly to the matter of field differences. These differences are muted in the Graduate Fellowship Program by the fact that selection is made within field, so that there is no direct competition of members of one field with those of another. In spite of the fact that these differences have not been a major focus of interest, they do show up from time to time in these studies, confirming the evidence from other researches on scientific manpower. Following graduation, field differences were also found with respect to the nature of the on-the-job functions of the scientists. Geographic differences, some of them correlated with employer categories, were found in the percentage of awardees and nonawardees who went to, or remained in, the various regions of the United States, as shown by their locations at the time of follow-up. Although these findings were scattered, some attention to them is probably relevant to a general review of the significance of fellowship selection and the evaluation of a fellowship program.

In ability profiles, some quite startling differences were found in the typical patterns of the several fields. The most striking contrast is that of the average ability pattern of engineers and the corresponding pattern of psychologists. The engineers are highest on a measure of quantitative ability and are relatively low on verbal ability. Exactly the reverse is true of the psychologists. In general level of the several ability measures, the physicists and mathematicians were found to be outstanding (10, p. 3); this confirms a finding from a study of the high school ability patterns of PhD's, most of whom had not been NSF fellowship candidates. The ability patterns found among the NSF candidates were stable from year to year, as were the high school ability patterns, thus indicating an ordered relationship which is both nationwide and durable.

Members of the various fields differ significantly in the amounts of time they devote to various functions on the job after graduation. It was found, for example, that many more chemists than members of other fields spend approximately 100 percent of their time in research activity. Physicists whose time is devoted exclusively to research are found in a somewhat smaller proportion than the chemists, but still far more frequently than is true in any other field. Fields with a high percentage of people who do little else but teach are mathematics, biology, psychology, and geology (24, pp. 23-24).

REGIONAL ADVANTAGES

Regions and employer categories may be compared with regard to the extent to which they are able to attract the more capable of the former fellowship candidates when these people graduate and go into regular jobs. For this purpose, it is assumed that, among the fellowship candidates, awardees as a group are more able than nonawardees. While certainly not uniformly valid with respect to individuals, this distinction has a useful degree of validity when applied to groups of dozens or hundreds. A given region of employer category will then be said to have an advantage if it attracts a higher percentage of the former candidates who have won awards than of those who have not. In these terms, academic and industrial establishments in New England had an advantage over those in other regions; in the Middle Atlantic area these groups were at a disadvantage. Government departments in both areas had an equitable

balance of awardees and nonawardees. In the Midwest academic institutions had an advantage; government suffered by comparison. In the South and West (except California) the situation was just the opposite of that in the Midwest; in California governmental organizations were at a disadvantage (20, p. 14). By themselves, these findings are a bit difficult to interpret. It may be noted, however, that in general, those regions which seem to have an advantage with respect to postdoctoral employment of awardees are the same regions which tend to produce a high proportion of all candidates. It is important in this connection to remember, too, that the candidate group as a whole, including nonawardees, is superior to the general run of graduate students (24, p. 40).

SOME EMPLOYERS MORE EQUAL

When the first postdoctoral employers of former fellowship candidates are compared with those of doctorate-holders in general, it is to be noted that the academic institutions, which in general get a better than equal share of former fellowship candidates, get an even greater proportion of awardees (24, p. 44). Among industrial and business employers there are significant variations in the percentage of awardees: electrical and electronic manufacturing have an advantage, followed by research institutes and foundations; pharmaceutical firms are about neutral in this respect, while aircraft manufacturing is at a disadvantage. Within government, the higher percentage of awardees was found in the Department of Agriculture. Next, there was a group of other civilian governmental agencies; the military departments were at a disadvantage. State governments employed none of the awardees and only a small percentage of nonawardees (20, p. 16).

OTHER FINDINGS

There are many findings that do not fit neatly under such headings as "reliability," "validity," and "field differences." Yet these findings, taken together, afford a number of clues as to the impact of a fellowship program on the educational process, and also serve to describe the process and its outcomes. A number of these findings were of immediate interest in the development of the fellowship evaluation operations, and served either to confirm procedures that were tried out tentatively or to suggest revisions of operating procedures. Some of the highlights of this group of miscellaneous findings follow, with interpretations of their significance where it is apparent.

INTERDISCIPLINARY PANELS

Multidiscipline panels were used successfully at the postdoctoral level from the inception of the NSF Fellowship Program (9). It seemed probable that such panels would be effective also in evaluating predoctoral candidates, and an experimental panel was tried out in 1955. It was composed of highly experienced panel members, including panel chairmen. It was found that these men could evaluate candidates in all fields without any observable bias for or against their own fields of specialization. Their judgments also agreed very well ($r = .78$) with the judgments rendered by the operational panels in the candidates' own fields. This gave encouragement to further experimentation on a broader scale and led to the assembling of all available evidence on the question of "field bias" in such panels. In all, eight experimental and operational situtations were investigated by analysis of variance, and in no case was there any evidence of "field bias" (14, p. 10). It was also the consensus of the participants in these studies that most of the evaluational process did not require special knowledge of a particular field. When such special knowledge was required, it could usually be obtained by informal consultation. One immediate result of this investigation was the abandonment of the process hitherto used in the field of biology, where several subpanels in the various biological specialties had been employed. A "bio-interdisciplinary" procedure was used subsequently, with a considerable reduction in operational complexity, and apparently with some improvement in over-all efficiency—certainly an improvement in flexibility of assignment of panel personnel.

One of the problems encountered by the panels in the use of reference reports was that some applications were submitted by students who had recently entered a new department. Could the reference reports be relied upon under these circumstances? It turned out, on research investigation, that there was no difference in average rating associated with length of residence in a department, nor were reliabilities of reference report ratings related to length of residence (14, p. 8). What apparently happened when students were not well known in their new departments was that they secured reference reports from their old departments in which they were known. No doubt, the problem of evaluating the significance of such reference reports remained for the panelists, but the reports per se were both reliable and unbiased in the general average of the ratings rendered.

INSTITUTIONAL VARIATIONS

One study turned attention to the institutions from which the fellowship candidates came, and to students from institutional groups. Here it was found that the higher the level of degrees granted by an

institution, the higher the average Advanced Test score of the students. That is, the average of the Advanced Test scores of students from schools that granted Master's degrees was higher than that of students from BA-only schools; on the average, students from schools that granted the PhD attained still higher scores.

A similar trend was noted with size of university. There was an interesting variation in average scores from institutions that granted Master's degrees. First-year students from these schools attained higher scores on the Advanced Test than did the intermediate candidates from the same schools. An interpretation could be that the better students from these schools went elsewhere for their graduate education; they were then recorded as coming from the new graduate schools. The poorer students, staying at their alma maters, or making up deficiencies before going on to stronger institutions, showed less knowledge of their fields, on the average, than did the seniors from the same institutions. Many of the latter went elsewhere for their graduate work. Whether this is the explanation for the finding could not be determined from the data of this first study (34, p. 2). It would be possible to make a better-controlled study from the more extensive records and data-processing systems now available. Another finding of this same study was the fact that even the best graduates of the teacher-training institutions seldom attained Advanced Test scores equal to the average of all candidates at the same level.

EMPLOYER CATEGORIES

The follow-up studies of fellowship candidates indicated that, 6 to 9 years after application, about half of the terminal candidates were in academic jobs, about a third of them in business and industry, and about one tenth in government positions. The first-year and intermediate candidates, followed up from 1 to 5 years after the doctorate, indicated a net shift away from academic employment, which most of them entered immediately after graduation, to the less frequent categories of business and government. For the awardees, where the tendency to stay in academe immediately after graduation was strongest, this later shift was also strongest. The net effect would be toward a more even distribution of awardee/nonawardee ratios among the three major employer categories (20, p. 9; 24, p. 43).

A number of findings were concerned primarily with matters of fellowship selection research design and procedures. While they are helpful in evaluating the significance of the results, their main merit is for the light they throw on the research process itself, and the comparisons and suggestions they may hold for others who might be involved in similar projects. Some of these findings in this category follow.

BIAS BY LEVEL?

The reference report ratings use as a normative base the general run of students at a given level. Accordingly, the distribution of scores should, in theory, be similar at all levels and should reflect an increasingly high standard as the less able students drop out. However, this is not always the case. In an early study (29, p. 6), the scores tended to go steadily upward, from 3.7 at Level 1 to 3.9 at Level 2 and 4.0 at Level 3, on a six-step scale whose scores ranged from zero to 6.0. What probably was happening was that the raters (mostly graduate professors) were adopting a rating standard

which tended to cut across all levels, and in which a student tended to maintain a constant rating score, rather than a variable one, to reflect the increasing competition. The variations here are small, however, and of no great significance. At the time this finding was made, the NSF program was not widely known, and possibly a larger proportion of the first-year students were coming from universities with graduate departments. At the present time, many first-year students come from 4-year colleges, and are rated in comparison with seniors, frequently by professors who are not in intimate touch with graduate school standards. A recent finding (25, p. 35) was that first-year students were rated about .25 point (one third of a standard deviation) above the intermediate candidates. As other evidence indicates that they are, on the average, slightly less able, this probably represents a minor amount of rating bias. In any event, however, this bias is small, and can be accommodated without distorting the selection process.

WHO RETURNS QUESTIONNAIRES?

Questionnaire return rates have uniformly been higher from those who were awarded fellowships than from those who were not. In the Atomic Energy Commission group, where an assiduous follow-up was made, the final return rates were 85 percent for nonawardees and 94 percent for awardees. In the Terminal Study, the rates were 71 percent and 85 percent, respectively (20, p. 3). It was also found that those who had attained doctorate degrees were much more likely to return questionnaires than those who had not graduated—the percentages ran about one third higher on the average (24, p. 13).

AND CONFIDENTIAL REPORTS?

Confidential Reports of Performance have been the main source of data regarding on-the-job performance. Yet this is a time-consuming method, both for the researcher and for those who are asked to complete the reports. Of the 6,000 Confidential Reports returned in the latest follow-up study, it was found that about three fourths were usable. That is, the raters had observed job behavior for a sufficient period of time, and sufficiently recently, with a good enough chance for pertinent observation, to lend credence to the ratings rendered. The other one fourth were deficient in one or more of these requirements. In this study, three or more Confidential Reports were secured for two thirds of the 1,619 individuals in the final validation sample; another one sixth had two reports; the remainder had only a single valid report.

REFERENCE REPORT DEVELOPMENT

The development of the reference report form was accomplished in part through extensive experimentation with variations in the report form during the first 2 years of the NSF Fellowship Program. The seven rating scales used in the 1952 reference report form (the term at that time was Confidential Report) were subjected to a factor analysis which showed that three factors served to describe all the significant variations in the scores. These factors were (a) ability to create a generally favorable impression, (b) ability to evaluate critically, and (c) dependability and self-reliance (2, p. 3). The 1953 reference report form was composed of a large number of rating scales.

These, together with the test scores (including an experimental test), the undergraduate grades (including an improvement-in-grades index), and Quality Group were also subjected to a factor analysis. The members of the group involved here were all first-year candidates; this is an important factor in interpreting results, as the ratings were necessarily based on observation of undergraduate, rather than graduate school performance. A total of eight factors were separated, but only six could be reasonably well identified; the remaining two were ill-defined. The six factors were described in terms of the set of scores on which they had major loadings. These were: (a) test scores, (b) a general rating factor, (c) undergraduate grades, (d) emotional stability, (e) ability to generate new ideas, and (f) career drive (3, pp. 14-17). Four of these were concerned primarily with the rating scales. Consideration of these factors was helpful in development of the next version of the reference report form. That form, worked out during the summer of 1954, has served, with only minor revisions, to the present time.

A DESCRIPTIVE REPORT SCORE

An attempt was made, in 1954 and 1955, to determine whether the free-response portions of a reference report form yielded scorable information that was not contained in the over-all rating. By "scorable" is meant here information that could be derived and stated quantitatively by an intelligent clerical worker reading the report with the intent of determining the confidence the rater had in the student's ability to successfully complete graduate school. It was found that such determinations could be reliably made by the clerical staff, but at a cost which would be prohibitive except for research work on a limited number of cases. It seemed probable, from the statistical evidence, that such scoring of the free-response comments did yield unique information, although this could not be determined unambiguously because of the unreliability of the individual ratings. A later examination of the validity of such "descriptive report scores," as they were called, indicated that they were slightly, but not significantly more valid for the prediction of a later on-the-job criterion than was the over-all graphic rating. The outcome of this attempt was negative as far as application to the evaluation process was concerned; it did tend to give statistical support to the opinion of the panelists that such comments add a significant element—that it is insufficient to note only the over-all rating scale.

CURVILINEAR RELATIONSHIPS

Turning to the more complex matter of on-the-job performance, statistical analysis provided a picture of the degree of intercorrelation among the many elements available for combination into a criterion score. The first study of this nature was done in a follow-up of the candidates for Atomic Energy Commission fellowships. This program, which immediately preceded that of the National Science Foundation, used very similar procedures; a group of 1948-1949 candidates of the AEC program provided the first information regarding the validity of the techniques used in the National Science Foundation Program. In that follow-up, four measures in addition to Confidential Report scores were obtained. These were (a) publications, (b) number of people supervised, (c) highest academic level of those supervised, and

(d) income. It was found that publications did not correlate highly with any of the other variables; number of publications increased as income went up to a maximum (at about the 90th percentile of income) and declined thereafter. A similar nonmonotonic relationship was found between number and level of people supervised (13, pp. 7-9). A plausible explanation is that supervisory duties occupy more of the scientists' time as their careers and incomes advance; at the same time, output of scientific papers tends to fall off. No factor analysis of the criterion variables was attempted in this study.

AWARDEES VERSUS NONAWARDEES

Another set of findings concerns what may be the effects of the fellowship program—or at least the observed differences between awardees and nonawardees several years subsequent to application. It may be that these differences are not due to the award of a fellowship, but result from the differences within the candidates at the time of application. Or it may be that the observed differences are due to a combination or interaction of superior ability and the advantages of holding an NSF fellowship. From the observed data in these cases, it is not possible now to determine how much of the gross difference in any case may be attributable to either source of variation. A brief summary of the more typical differences follows.

DIFFERENCES GALORE

Among the 1949 AEC candidates followed up in 1955 it was found that, although the awardees were younger at the time of follow-up and had fewer dependents on the average, a higher percentage were married. A higher percentage had PhD's—a finding that has recurred consistently in all follow-up studies (30, p. 2). Among the terminal candidates, awardees in academic settings were doing more research; nonawardees were doing more teaching (20, p. 20). Of the awardees, 23 percent had had post-doctoral training, as compared with only 15 percent of the nonawardees (20, p. 13). There was little income difference between the terminal awardees and nonawardees; among the 1955-1956 first-year and intermediate candidates, the awardees had higher incomes (24, p. 31). Awardees publish more than do nonawardees (20, p. 49) and report greater career satisfaction and more progress toward career goals (20, p. 36). Awardees consistently get more peer nominations as outstanding scientists, and are more frequently cited in spontaneous comments as being outstanding in one way or another (22, p. 41). They get more academic honors, but the nonawardees are likely to report more honors later—holding of offices in scientific societies, for example (20, pp. 49-52). Awardees are more likely to have what might be considered a "good balance" of research and teaching—somewhat more than half time spent in research, but with a fair teaching load also, and perhaps some administrative responsibilities (24, p. 23). Nonawardees are more likely to have either heavy teaching loads, or research exclusively. On the average, the awardees are about a year and a half younger than the nonawardees at the time of the PhD, having spent 1 year less in graduate school (24, p. 37). Finally, when the two groups are compared on a rather comprehensive criterion of on-the-job achievement, the awardees are about two thirds of a standard deviation ahead of the nonawardees (26, p. 37). Yet even the latter, it is to be noted, have achieved academically considerably more than PhD's in general, being younger by about 2 years, and having spent a year and a half less in graduate school (24, p. 37).

A CONTEXT FOR INTERPRETATIONS AND CONCLUSIONS

PLURALISM IN SUPPORT

The growth of graduate education and the increase in direct support of graduate students through fellowship stipends over the past decade and a half have greatly increased the importance attaching to methods of selection of those who will obtain fellowships. The number and variety of programs, by several government agencies as well as by various private organizations and by the universities themselves, insures that a large proportion of doctoral candidates, particularly in the sciences, will be able to obtain a sufficient subsidy to keep them in graduate school on a full-time basis. The impact of a program devoted to awarding fellowships on the basis of ability alone is different in the context of such a variety of modes of support from what it would be if this kind of program were the only important source of fellowship support. The significance of such a program emphasizes the particular significance of the way in which it is carried out.

A WAVE IS FELT AT THE MARGIN OF THE POOL

When the great majority of students have stipend support, and such support may come from many different sources, the amount of money put into the pot by any single source has a marginal effect. To resort to an oversimplified model of the situation for diagrammatic reasons, let us assume that, at a given point in time, there are 5,000 PhD candidates in the sciences, and that there is fellowship support for 3,000 of these and assistantship support for 1,000 additional people, leaving only 1,000 "self-supported." The effect of adding funds to support 500 more Fellows would be to move 500 of the self-supported people to fellowship support. If all fellowship programs were run on "merit" bases, it would be assumed that the 1,000 self-supported students would, in general, be marginal ones. The net effect, then, would be that the new program would support 500 marginal students. This would be true regardless of the merit limitations put on the selection of the 500 new Fellows, even if those particular people should be the very top 10 percent out of the 5,000 total. The funds released by putting them on the new program would be shifted to others, and, through devious adjustments that cannot be followed in detail, finally 500 marginal students would get support from marginal programs—either fellowships or assistantships.

INPUTS ESTABLISH GRADIENTS

The importance of new money in fellowships could readily be underestimated from this oversimplified model were the dynamic aspects of the graduate school scene—the interplay of the various forces—overlooked. To oversimplify again, it should be noted that a new program of high prestige and selectivity tends to shift the attention of students, faculty, and administrators somewhat, to orient them in the direction of the value schema involved in the new program. If selections are made rigorously on merit, applicants will strive for merit in whatever terms that concept is defined in the program. The major effect, then, of such a high-prestige program of fellowships eagerly sought is to modify the value-structure and the direction of attention of the people involved. On the other hand, if an equal amount of money were to be made available without any merit requirements, it would tend to be used for the support of those who are unable to qualify for the

more meritorious awards. A heavy influx of such "low-merit" funds would therefore tend to distract attention from the "merit" programs, and to undercut the value system of the "merit" programs. The striving for excellence (however defined) would be watered down. The important point of this observation is that the same amount of fellowship money is involved in either case, and probably (considering the system as a whole, all programs combined) the same people are supported. But in the one case, the new and growing tip of fellowship support is at the high-quality end of the spectrum, and in the other case at the low-quality end. In the one case, the field is structured positively, in the other negatively by the same amount of funds, even though the same individuals are, overall, supported by fellowship funds.

In the context of this view of the effect of fellowship programs, the experience of a program of high-merit fellowship funds, constantly expanded over a period of 14 years and based upon merit selection, is particularly significant. That is the context in which this report of fellowship selection research occurs. The success of the NSF program in meeting the requirement for selection on the basis of ability only, and the various efforts made to improve selection techniques thus have significance far beyond the program itself. The experience of this research effort is worthy of particular attention by those concerned with other programs to improve the quality of graduate education. The experience of the NSF, the limitations of the NSF experience, and the problems remaining unsolved within the NSF program are of particular significance for other organizations which seek to improve their selection techniques. Much of what was learned here may be directly applicable in other governmental agencies should they choose to employ similar techniques. The fellowship programs of universities or private foundations are in many cases able to go beyond the experience of the NSF program in terms of the objectives of their selection processes, in terms of the research techniques which they might use, and, finally, in terms of the selection instruments which would be feasible to employ. Private auspices on the one hand, and local rather than national geographic scope on the other, provide advantages as well as disadvantages; perhaps some of the techniques suggested for the NSF program, which have never in fact been employed there, might find use in some of these other programs.

WHITHER?

A review such as this is incomplete without a look at the work that might have been done but was not, and at problems remaining to be solved. Some insight into the former may be gained by examination of the minutes recording the activities of the Office of Scientific Personnel Advisory Committee on Fellowship Selection Techniques. Another line of evidence is derived from a series of conferences called to consider selection of Fellows, and from conferences concerned with identification of creative scientific talent. The latter, sponsored by the National Science Foundation, were quite independent of the fellowship program as such, but were attended by the Director of Research of the Office of Scientific Personnel, because their outcomes had promise of application in fellowship selection.

A DECADE OF RECOMMENDATIONS

From the very first meeting of the Committee on Fellowship Selection Techniques in 1952, to its last meeting before being broadened to the Research Advisory Committee, the question of new selection devices came up for discussion. Every meeting either considered some specific technique, reviewed evidence with regard to some experimental work done elsewhere, or recommended a specific line of research in the area of noncognitive traits. These deliberations, spread over a whole decade, and including an almost complete turnover of committee personnel, never failed to stress the importance of branching out into new types of measurement that might tap additional valid variance. These deliberations brought to the fore the fact that there did not exist at the time any specific tested instrument which could be employed without further experimental tryout. They stressed the need for experimentation, either directly on candidates for fellowships or through university researchers who might employ samples of graduate students similar to those who apply for NSF fellowships. Throughout these meetings, also, the point was reiterated that such experimental predictors should be validated against on-the-job criteria. The point was also made that it would be impossible to be sure of the utility of a technique unless its results were set aside, perhaps for several years without direct use in selection, until suitable crtierion data might mature. The variables derived from the new technique could then be validated against on-the-job criteria. In brief, the committee recognized from the beginning the long-term nature of such research.

CRITERIA AND PREDICTORS TO BE TESTED

Similar recommendations came from a series of conferences, some convened directly by the National Science Foundation, some sponsored by others and supported by the NSF. One of the first conferences to recommend exploration of the use of noncognitive measures took place in November 1953 at the National Academy of Sciences—National Research Council. It was held primarily for the purpose of defining reasonable criteria of success in science (4), yet the conference also considered predictors of such success, and the late L. L. Thurstone specifically recommended that investigations be extended beyond the realm of cognitive abilities. He suggested that further significant improvement might be made by the employment of a number of objective laboratory procedures that he briefly described. He suggested, also, as did others at the conference, the use of

biographical information blanks which might identify features of an individual's life history and experience which would be predictive of later success in science.

CAN WE MEASURE CREATIVITY?

Another conference, in June 1954, considered techniques of selecting Fellows. This conference was called by the National Science Foundation and held in its offices. The conference concerned itself with the definition of "ability" as required by the Act setting up the NSF, and various measures of scientific ability. Among the abilities considered was "creativity." It was the consensus of the meeting that, important as creativity is, no adequate predictors of this quality are currently available. It was recommended that research in this area be supported, in the hope that adequate measures might eventually become available. Partly as an outcome of this recommendation, the NSF supported a series of conferences, sponsored by the University of Utah and led by Dr. Calvin W. Taylor, on the identification of creative scientific ability. These conferences, held at mountain retreats near Salt Lake City, have resulted in a series of reports and two hardback books on the subject of research on creativity.

TRY BIOGRAPHICAL DATA?

In August 1956, the National Science Foundation sponsored a conference in Chicago (chaired by Dr. Kenneth E. Clark, then of the University of Minnesota) on "The Use of Objective Tests in the Selection of NSF Fellows." The conference endorsed procedures then in use, recommended that the NSF maintain a search for potential new selection instruments, and conduct experimental work on the long-term validity of such instruments. One of the specific items discussed was the use of a biographical information blank.

Despite these recommendations, experimentation on applicants for fellowships was deemed impractical because of the constant possibility of political repercussions. Even if the data from experimental instruments were not made available to the panels and thus could not be used in selection, it was feared that misunderstandings might develop, requiring defense of instruments which were outside the usual definitions of "ability." The decision of the Foundation was that it should not risk the possible consequences of such misunderstandings, and the work was never funded. Meantime, one opportunity did develop where one of a large number of possible instruments could be given a tentative check-out as to its long-term validity.

NO HITS, NO RUNS, ONE ERROR

At the 1962 Utah Creativity Conference, a test was described that appeared to have good potential for measuring some aspects of ability not encompassed by the instruments in current use. This was the Remote Associations Test of Dr. Sarnoff Mednick, of the University of Michigan. It was possible to perform a minor informal experiment on this test by mailing it to a sample of former fellowship candidates on whom criterion data were already available. These people cooperated very well, completing the test and returning it for scoring by Dr. Mednick. In brief, it was found that the test had little validity for predicting the kind of criterion data developed in the Office of Scientific Personnel follow-up studies. It seemed to measure some type of verbal ability, correlating modestly with the GRE verbal score.

THE QUESTION REMAINS

Many other potential instruments have been developed over the past decade or so; some of them have been tested on populations of college and graduate students. None has yet been tested under conditions which would permit a clear statement of its validity for fellowship selection, above and beyond the validity of the procedures in present use. Perhaps the major reason why such tests have not been conducted is the fact that so many years must elapse between administration of the experimental instrument and a test of its validity. Few researchers in universities are in a position to undertake a study of such duration. Yet the question remains: could not the present techniques be supplemented in such a way as to result in a marked rise in validity if measurements in the area of noncognitive traits were to be added to the present battery?

REFERENCES

1. An Analysis of the Evaluation Panel Suggestions on the 1953 NSF Fellowship Material, by Calvin W. Taylor, 25 February 1954. 13 pp., mimeographed.
2. A Factorial Study of the 1952 NSF Confidential Report Ratings, by Calvin W. Taylor, 30 March 1954. 8 pp., mimeographed.
3. A Factorial Study of 28 Predictor Scores in the 1953 NSF Fellowship Program, by Calvin W. Taylor, 5 April 1954. 25 pp., mimeographed.
4. Discussions on "Criteria of Success in Science." Report of a conference held in November, 1953, assembled by Calvin W. Taylor, 15 April 1954. 55 pp., mimeographed.
5. An Analysis of the Evaluation Panel Suggestions on the 1954 NSF Fellowship Materials, by Calvin W. Taylor, 22 April 1954. 12 pp., mimeographed.
6. A Description of Two Groups of Scientists on a Scientific Adviser Report Form, by Calvin W. Taylor, 17 May 1954. 10 pp., mimeographed.
7. Significance of Descriptive Comments on Confidential Reports, by Lindsey R. Harmon, 12 August 1955. 10 pp., mimeographed.
8. Evaluation of NSF Graduate Fellowship Candidates by an Inter-Disciplinary Panel, by L. R. Harmon, 15 August 1955. 7 pp., mimeographed.
9. The Selection of Postdoctoral Fellows by an Inter-Disciplinary Board, by L. R. Harmon, 15 August 1955. 9 pp., mimeographed.
10. Ability Patterns in Seven Science Fields, by Lindsey R. Harmon, 17 October 1955. 14 pp., mimeographed.
11. Are Confidential Reports Affected by Candidates' Length of Residence in a Department? by Lindsey R. Harmon, 25 January 1956. 9 pp., mimeographed.
12. "Field Bias" in Interdisciplinary Panel Evaluations, by Lindsey R. Harmon, 31 December 1957. 11 pp., mimeographed.
13. A Follow-Up Study of A.E.C. Fellowship Candidates, by Lindsey R. Harmon, 13 January 1958. 14 pp., mimeographed.
14. The Effect of Fellowships on Acceleration of PhD Attainment, by Lindsey R. Harmon, 15 January 1959. 13 pp., multilithed.
15. Validation of Fellowship Selection Instruments Against a Provisional Criterion of Scientific Accomplishment, by Lindsey R. Harmon, 15 December 1959. 10 pp., multilithed.
16. Effects of a Summary Score on Panel Judgments, by Herbert Soldz, 29 December 1959. 13 pp., multilithed.
17. Comparison of 1960 Cooperative and Regular Graduate NSF Candidates and Awardees, by Herbert Soldz, 20 September 1960. 32 pp., multilithed.
18. A Study of Graduate Fellowship Applicants in Terms of PhD Attainment, by John A. Creager, 22 March 1961. 41 pp., multilithed.
19. Development of Evaluation Composites in Graduate Fellowship Applicant Groups, by John A. Creager, 2 July 1962. 24 pp., multilithed.
20. Some Characteristics of Former Fellowship Applicants Six to Nine Years Later, by John A. Creager, 15 May 1962. 90 pp., multilithed.
21. Regression Analyses in the Cooperative Graduate Fellowship Program, by John A. Creager, 25 June 1962. 16 pp., multilithed.
22. The Terminal Follow-Up Validation Study, by John A. Creater, 15 December 1962. 16 pp., multilithed.
23. Formation of Criterion Composites for Predictor Validation, by John A. Creager, 20 May 1963. 20 pp., multilithed.
24. Some Characteristics of First-Year and Intermediate Fellowship Applicants 8 to 10 Years Later, by John A. Creager, 16 August 1965. 56 pp., multilithed.
25. Predicting Doctorate Attainment with GRE and Other Variables, by John A. Creager, 16 November 1965. 48 pp., multilithed.
26. On-The-Job Validation of Selection Variables, by John A. Creager and Lindsey R. Harmon, 15 April 1966. 70 pp., multilithed.
27. Comparison of Subjective Judgments of NSF Applicants Who Applied in Each of Two Successive Years. Appendix 1 to Third Research Progress Report, by Calvin W. Taylor, 30 June 1954.
28. History and Rationale Underlying the 1954 NSF Confidential Report Form. Appendix 2 to Third Research Progress Report, by Calvin W. Taylor, 30 June 1954.

29. Relationship of 4 Panel Roster Scores to Quality Group Evaluations of NSF Graduate Applicants. Appendix 3 to Third Research Progress Report, by Calvin W. Taylor, 30 June 1954.
30. Research Progress Report, FY 1955. Fellowship Research: Procedures and Results, by Lindsey R. Harmon, September, 1955. 10 pp., mimeographed.
31. Fellowship Selection Research: A Four-Year Progress Report. NAS-NRC Publication 564, by Lindsey R. Harmon, January, 1958. 37 pp., mimeographed.
32. The Summary Score in the Graduate Fellowship Program. (For Official Use Only.) No author listed. Printed 6 September 1961. 5 pp., multilithed.
33. Fellowship Research Working Paper: Effects of Prescreening Graduate Fellowship Applicants, by John A. Creager, 9 February 1962. 4 pp., multilithed.
34. Special Report #1. Ability Patterns of Institutional Types (Privileged Document), by Lindsey R. Harmon, 29 December 1955. 7 pp., mimeographed.
35. Special Report #2. A Comparison of the 1959 Candidates and Awardees in the NSF Regular and Cooperative Fellowship Programs (Privileged Document), by Lindsey R. Harmon, 20 January 1960. 8 pp., multilithed.

INDEX